Bayes-Klassifikator

Vorlesung 6, Maschinelles Lernen

Dozenten: Prof. Dr. M. O. Franz, Prof. Dr. O. Dürr

HTWG Konstanz, Fakultät für Informatik

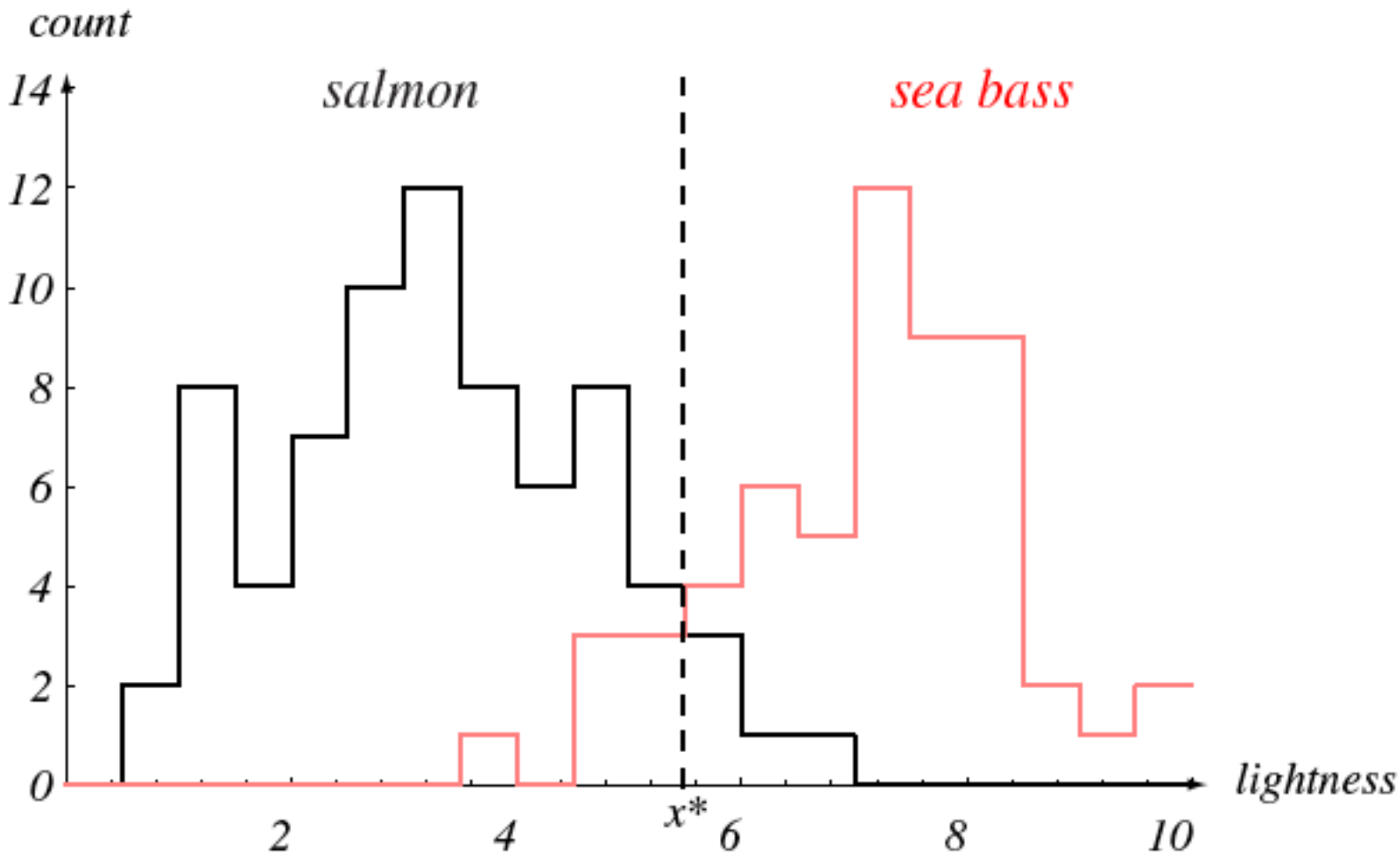
Übersicht

Übersicht

Entscheidungstheorie

- Die Entscheidungstheorie bewertet die Kosten einer Klassifikationsentscheidung mit Hilfe von Wahrscheinlichkeitsaussagen.
- Grundannahme: **alle Wahrscheinlichkeiten seien im vornhinein bekannt**.
- Bildet die formale Grundlage der Mustererkennung.
- Formalisiert den “gesunden Menschenverstand”.

Beispiel für eine Mustererkennungsaufgabe



[Duda et al., 2001]

Entscheidungstheoretischer Ansatz

- Annahme: Die Reihenfolge der Fische sei zufällig, d.h. nicht vorhersagbar.
- In der Entscheidungstheorie heißt das: die Natur ist in einem von zwei möglichen Zuständen ω (”state of nature”):

$$\begin{aligned} \text{Lachs:} \quad & \omega = \omega_1 \\ \text{Wolfsbarsch:} \quad & \omega = \omega_2 \end{aligned}$$

  Weil ω unvorhersagbar ist, muss ω als Zufallsvariable beschrieben werden.

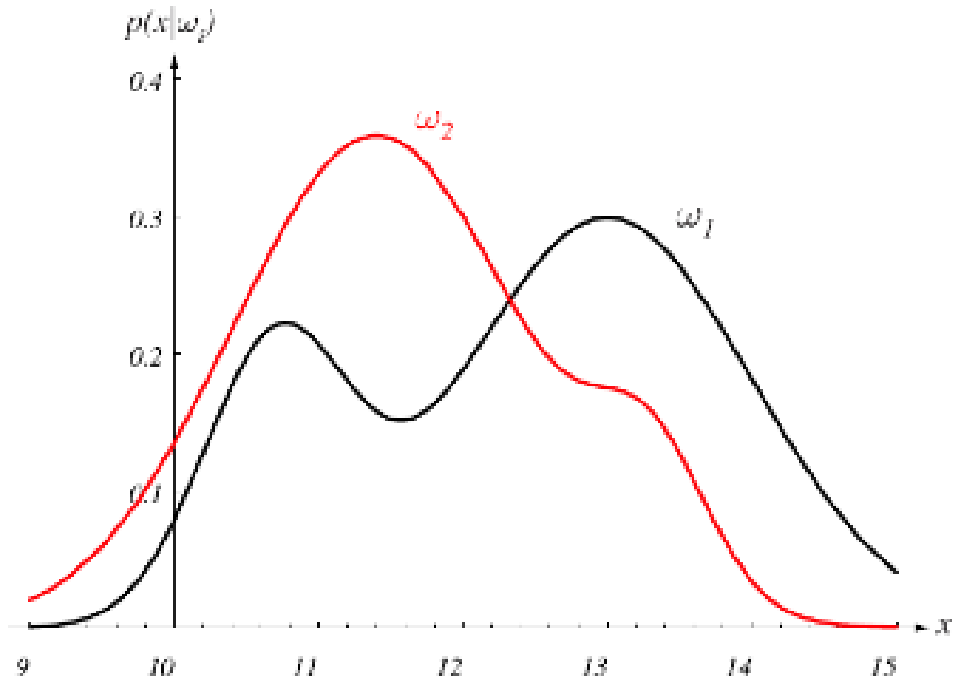
- Annahme: Es gebe eine A-priori-Wahrscheinlichkeit $p(\omega_1)$ für Lachs und $p(\omega_2)$ für Wolfsbarsch. Bei zwei Zuständen gilt

$$p(\omega_1) + p(\omega_2) = 1.$$

  Die A-priori-Wahrscheinlichkeit beschreibt das Vorwissen über das Entscheidungsproblem.

Klassenbedingte Wahrscheinlichkeit

- Wenn jeder Fehler gleich viel kostet, und wenn das Band nicht beobachtet werden kann, ist die beste Entscheidungsregel: "Entscheide für ω_1, falls $p(\omega_1) > p(\omega_2)$, sonst für ω_2."
- Gibt es (fehlerbehaftete) Messungen eines Merkmalvektors $\mathbf{x}$, dann hängt die Messung über die klassenbedingte Wahrscheinlichkeit $p(\mathbf{x}|\omega_i)$ von ω_i ab.



Paare von diskreten Zufallsvariablen

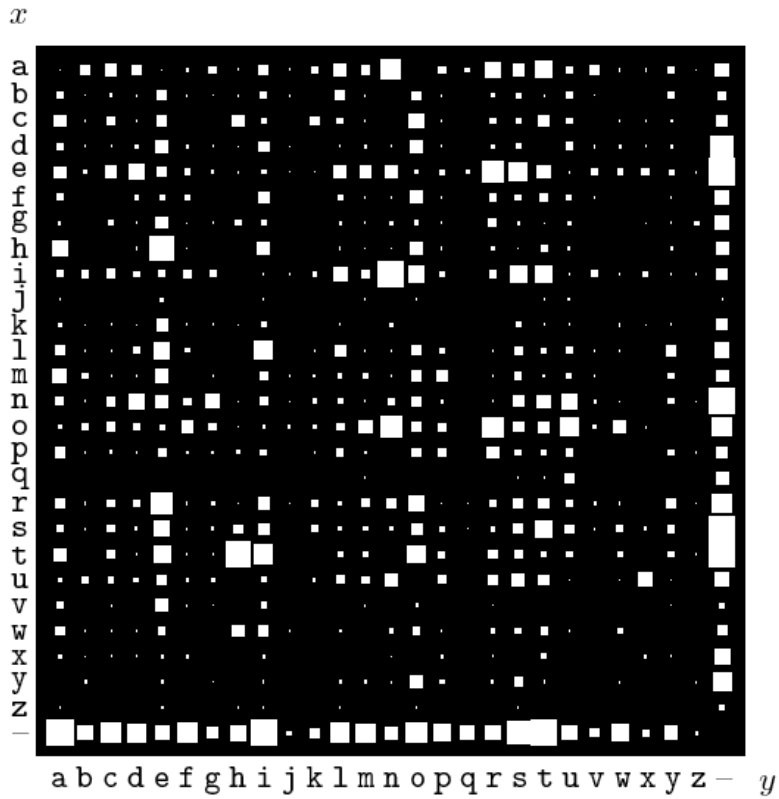
Zwei Zufallsvariablen x und y werden durch eine gemeinsame Verteilung $p(x, y)$ beschrieben, die jedem Wertepaar (x_i, y_j) eine Wahrscheinlichkeit p_{ij} zuordnet. Es gilt

$$p(x, y) \geq 0 \quad \text{und} \quad \sum_{x_i \in \mathcal{X}} \sum_{y_j \in \mathcal{Y}} p(x, y) = 1$$

Randwahrscheinlichkeitsfunktionen von x und y:

$$p(x) = \sum_{y_j \in \mathcal{Y}} p(x, y)$$

$$p(y) = \sum_{x_i \in \mathcal{X}} p(x, y)$$



[MacKay, 2003]

Bedingte Wahrscheinlichkeiten

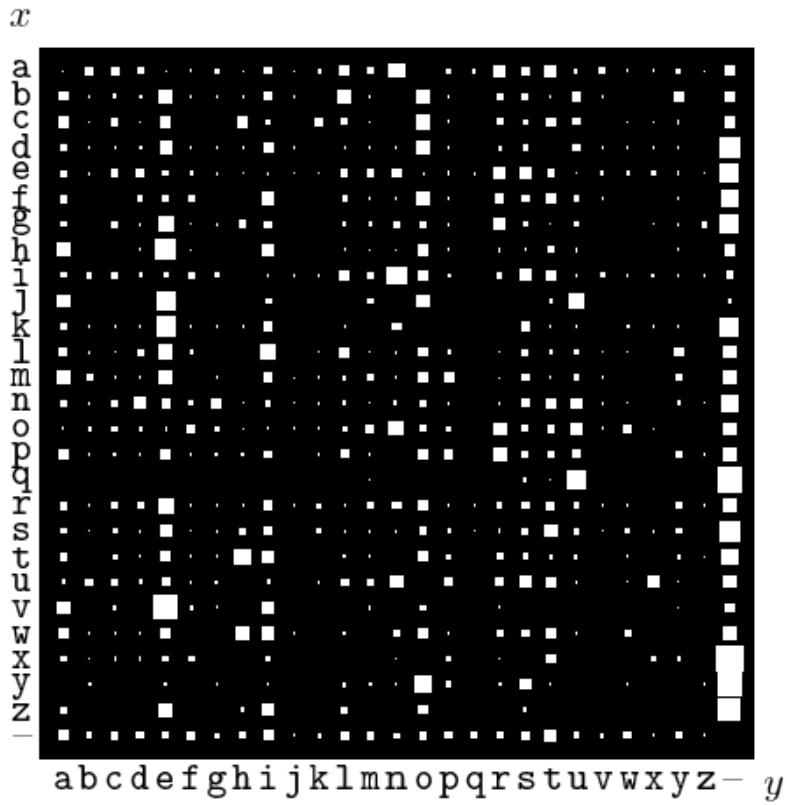
Wenn zwei ZV voneinander statistisch abhängig sind, erhöht Kenntnis der einen ZV das Wissen über die andere:

$$p(x|y) = \frac{p(x,y)}{p(y)} \quad \text{bzw.}$$

$$p(x,y) = p(x|y)p(y) \quad \text{(Kettenregel)}$$

$p(x|y)$: Wahrscheinlichkeit von x gegeben y.
Bei statistisch unabhängigen ZV gilt

$$p(x|y) = \frac{p(x)p(y)}{p(y)} = p(x)$$



(a) $P(y \mid x)$

[MacKay, 2003]

Der Satz von Bayes

Wie beeinflusst die Messung $\mathbf{x}$ die Schätzung des Naturzustandes?

Es gelten

$$p(\mathbf{x}) = \sum_i p(\omega_i, \mathbf{x}) \quad \text{(Randwahrscheinlichkeit von } \mathbf{x}\text{)}$$

$$p(\omega_i, \mathbf{x}) = p(\mathbf{x}|\omega_i)p(\omega_i) \quad \text{und} \quad p(\omega_i, \mathbf{x}) = p(\omega_i|\mathbf{x})p(\mathbf{x}) \quad \text{(Kettenregel)}$$

Substitution ergibt den

Satz von Bayes:

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{\sum_i p(\mathbf{x}|\omega_i)p(\omega_i)}$$

In Worten:

$$\text{A-posteriori} = \frac{\text{Likelihood} \times \text{A-priori}}{\text{Evidenz}}$$

Bayessche Entscheidungsregel

Nach einer Messung $\mathbf{x}$ ist die Fehlerwahrscheinlichkeit

$$p(\text{Fehler}|\mathbf{x}) = \begin{cases} p(\omega_2|\mathbf{x}) & \text{bei Entscheidung für } \omega_1 \\ p(\omega_1|\mathbf{x}) & \text{bei Entscheidung für } \omega_2. \end{cases}$$

Die Fehlerwahrscheinlichkeit für eine gegebene Messung $\mathbf{x}$ wird also minimiert durch die

Bayessche Entscheidungsregel

Entscheide für ω_1, wenn $p(\omega_1|\mathbf{x}) > p(\omega_2|\mathbf{x})$, sonst für ω_2.

bzw. entscheide für ω_1, wenn $p(\mathbf{x}|\omega_1)p(\omega_1) > p(\mathbf{x}|\omega_2)p(\omega_2)$, sonst für ω_2.

Damit minimiert die Bayessche Entscheidungsregel auch die Fehlerwahrscheinlichkeit über einen gegebenen Trainings- oder Testdatensatz.

Übersicht

Bayes-Klassifikator

Für einen gegebenen Meßwert x schätzt der Bayes-Klassifikator die Klassenzugehörigkeit nach der

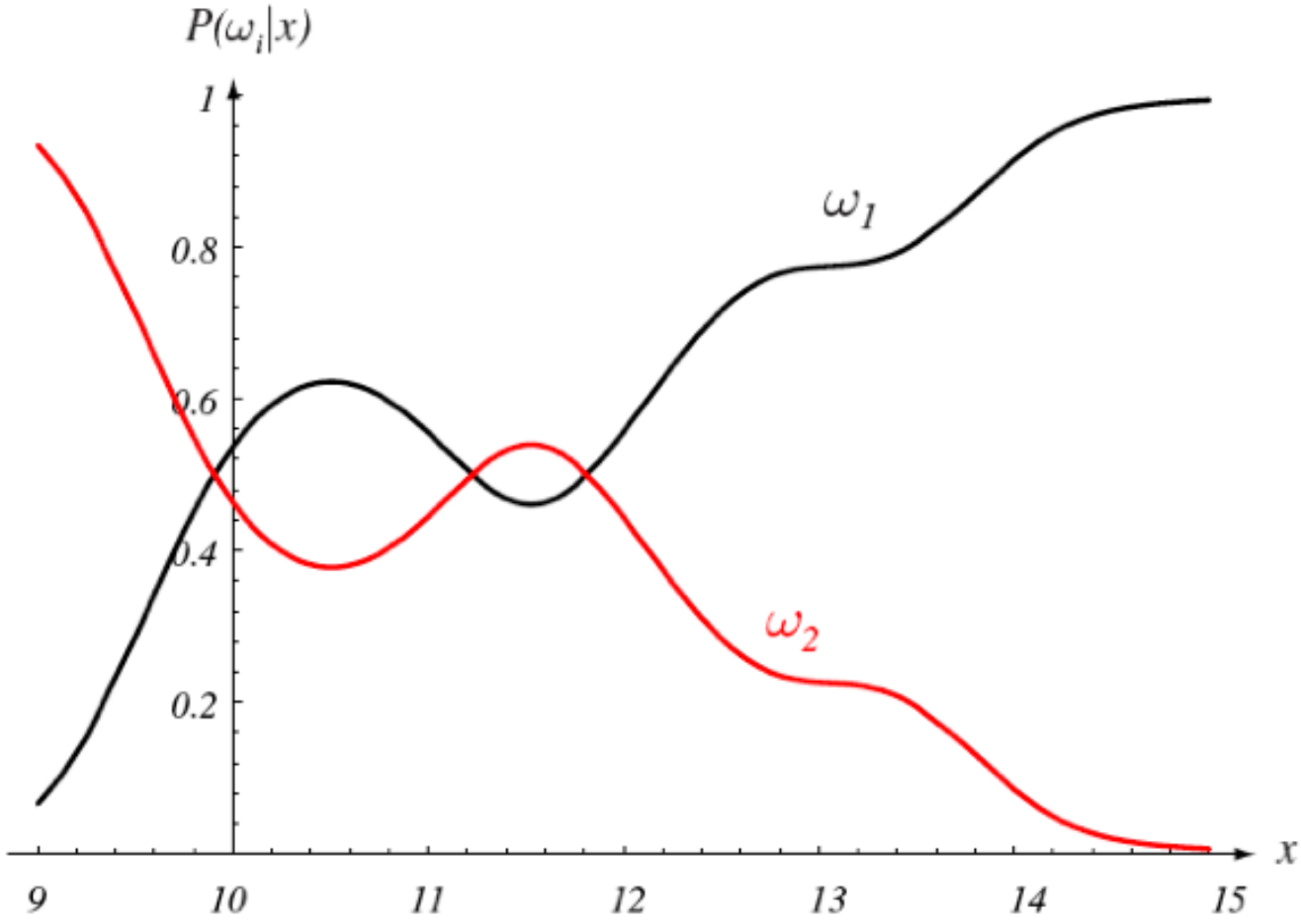
Bayessche Entscheidungsregel

Entscheide Dich immer für die Klasse ω_i, deren A-posteriori-Wahrscheinlichkeit $p(\omega_i|x)$ am höchsten ist.

Der Bayes-Klassifikator ist (theoretisch) der bestmögliche Klassifikator, da er die Wahrscheinlichkeit einer Fehlklassifikation minimiert. Jeder andere Klassifikator macht mehr oder mindestens gleich viele Fehler.

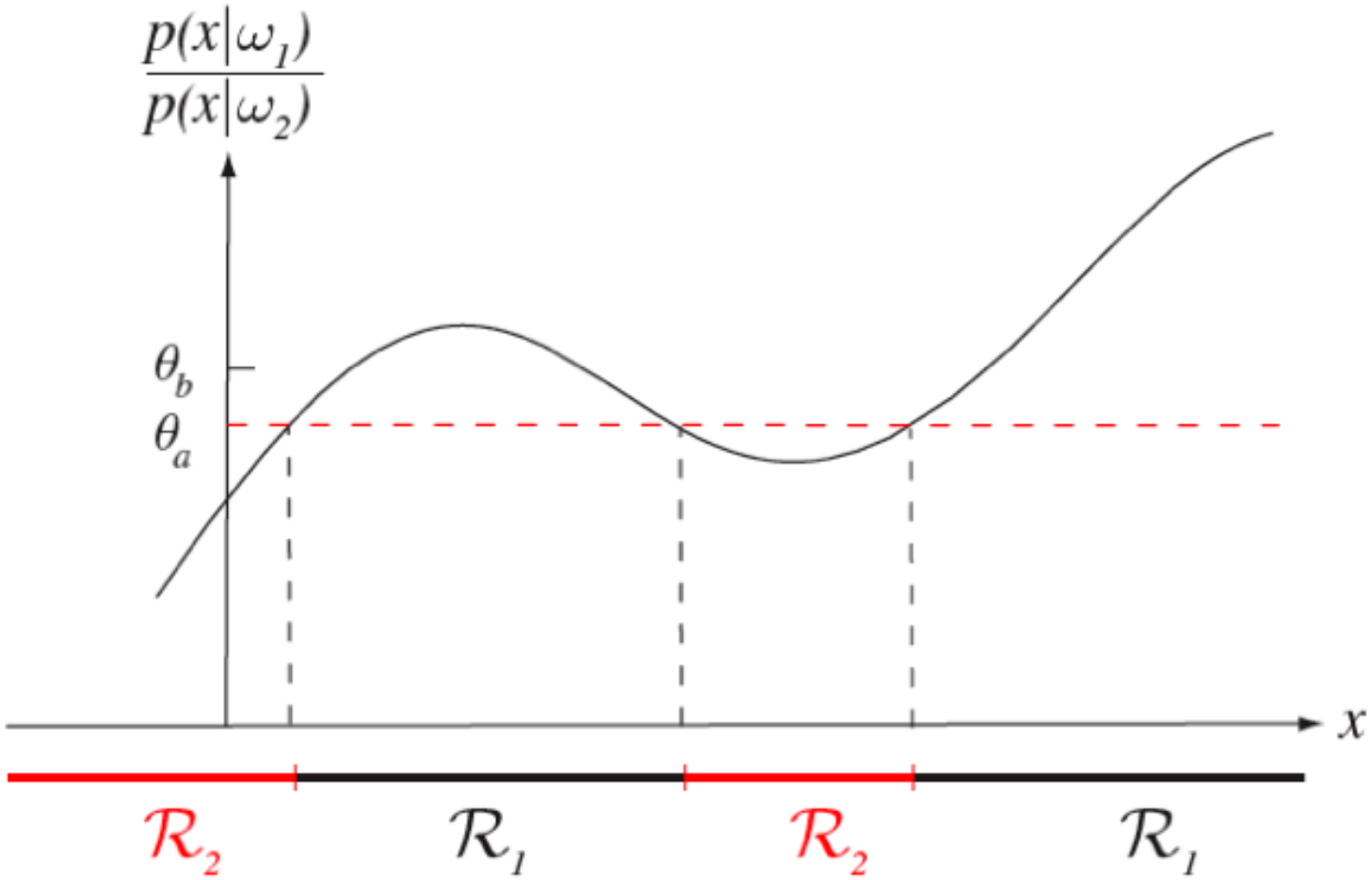
Je nachdem, welche A-posteriori-Wahrscheinlichkeit am größten ist, wird der Inputraum in Entscheidungsregionen aufgeteilt. Die Entscheidungsregionen sind durch Entscheidungsgrenzen voneinander getrennt.

Beispiel: A-posteriori-Wahrscheinlichkeiten für einen kontinuierlichen Messwert bei einem Zweiklassenproblem



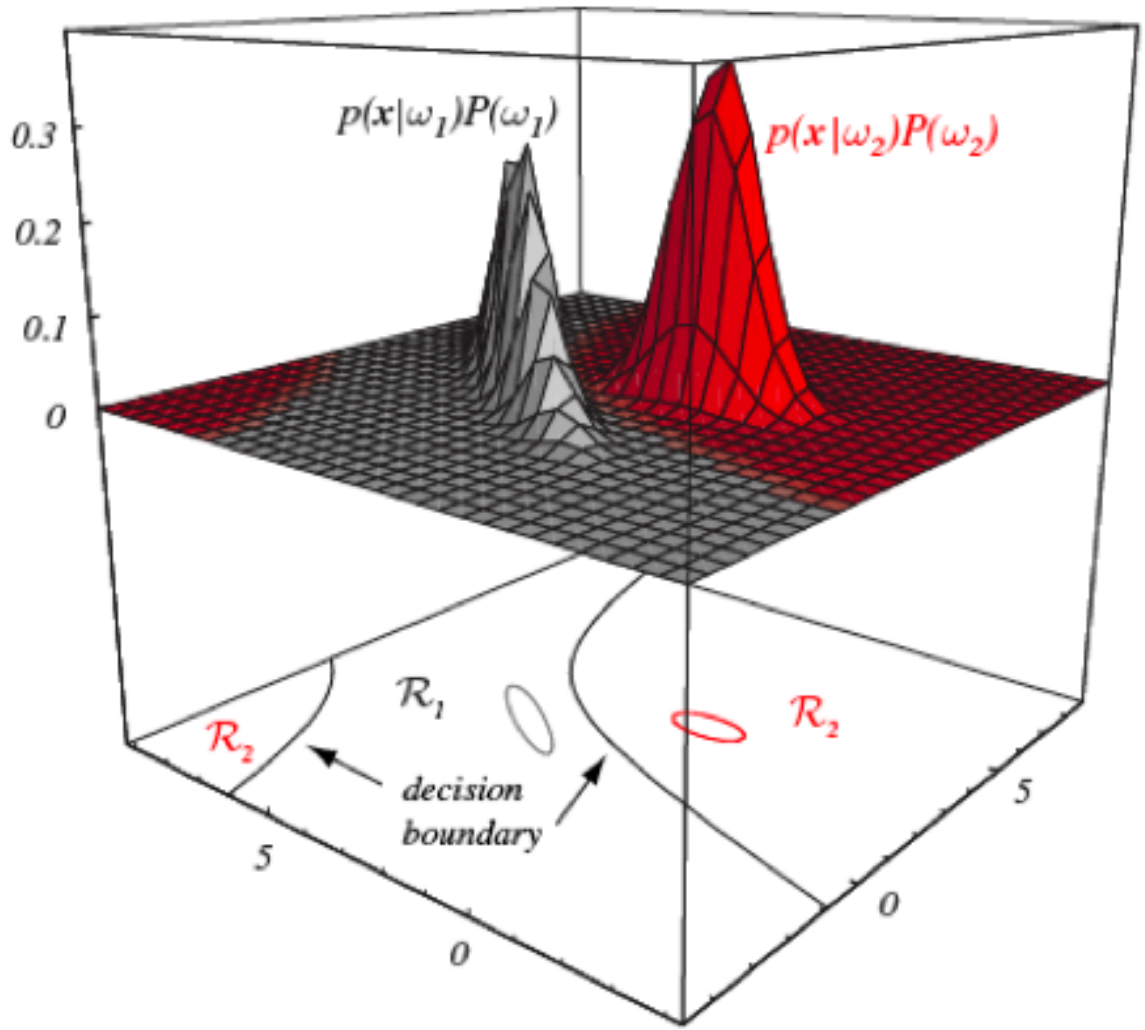
[Duda et al., 2001]

Beispiel: Entscheidungsregionen und -grenzen für einen kontinuierlichen Messwert bei einem Zweiklassenproblem



[Duda et al., 2001]

Beispiel: Entscheidungsregionen und -grenzen für zwei kontinuierliche Messwerte



[Duda et al., 2001]

Naiver Bayes-Klassifikator

Vereinfachende Annahme: die klassenbedingten Wahrscheinlichkeit bzw. Likelihood faktorisiert für die einzelnen Merkmale x_i im Merkmalsvektor $\mathbf{x}$, d.h. bei gegebener Klasse ω_i sind die x_i voneinander statistisch unabhängig:

$$p(\mathbf{x}|\omega_i) = p(x_1|\omega_i) \cdot p(x_2|\omega_i) \cdots = \prod_j p(x_j|\omega_i)$$

Die Bayessche Entscheidungsregel wird in diesem Fall zu:

Entscheidungsregel Naïve Bayes

Entscheide für ω_1, wenn $\prod_j p(x_j|\omega_1)p(\omega_1) > \prod_j p(x_j|\omega_2)p(\omega_2)$, sonst für ω_2.

Oft wird für die einzelnen Likelihoods $p(x_j|\omega_i)$ eine Gaußverteilung angenommen, man spricht dann von **Gaussian Naïve Bayes** (GNB).

Gaussian Naïve Bayes - Training

Beim GNB nehmen wir für jedes Merkmal eine Gaußverteilung an:

$$p(x_j|\omega_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{1}{2}\frac{(x-\mu_{ij})^2}{\sigma_{ij}^2}}$$

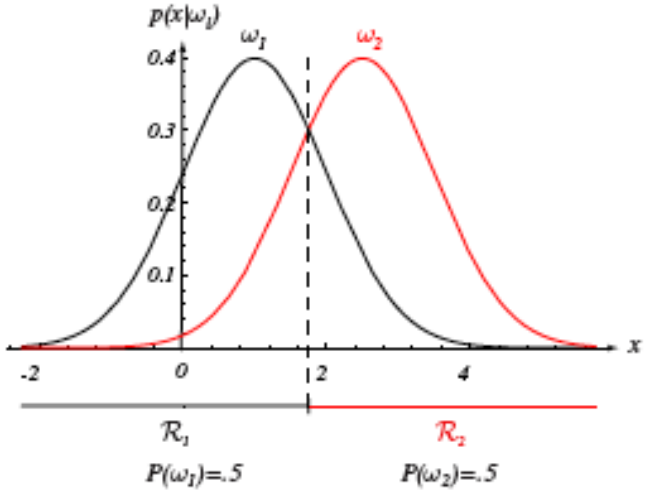
Der jeweilige Mittelwert μ_{ij} und die Standardabweichung σ_{ij} für jede Likelihood werden aus den Trainingsdaten $x_{j,k}$ geschätzt:

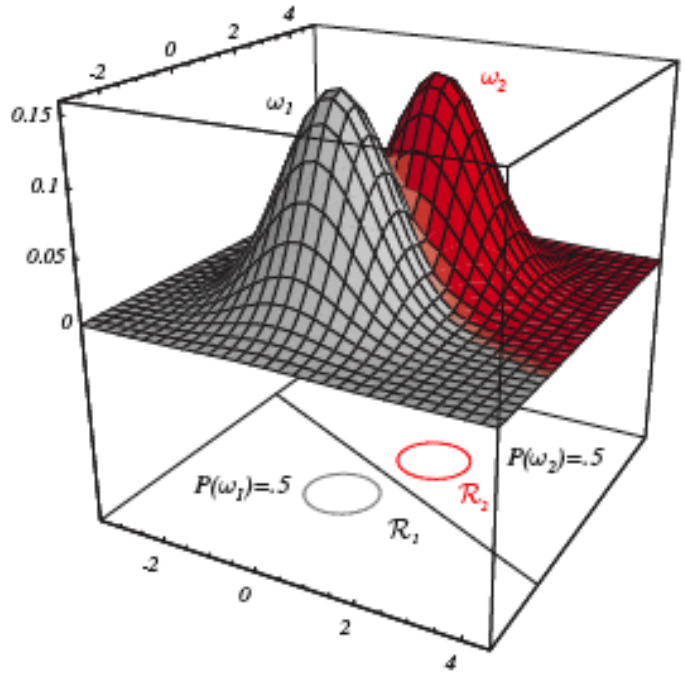
$$\mu_{ij} = \frac{1}{|\omega_i|} \sum_{k \in \omega_i} x_{j,k} \quad \text{und} \quad \sigma_{ij}^2 = \frac{1}{|\omega_i| - 1} \sum_{k \in \omega_i} (x_{j,k} - \mu_{ij})^2$$

Die A-priori-Wahrscheinlichkeiten für beide Klassen ergeben sich aus den relativen Häufigkeiten beider Klassen im Trainingsdatensatz:

$$p(\omega_i) = \frac{|\omega_i|}{\ell}$$

Beispiel: gleiche Varianzen in allen Merkmalen





[Duda et al., 2001]

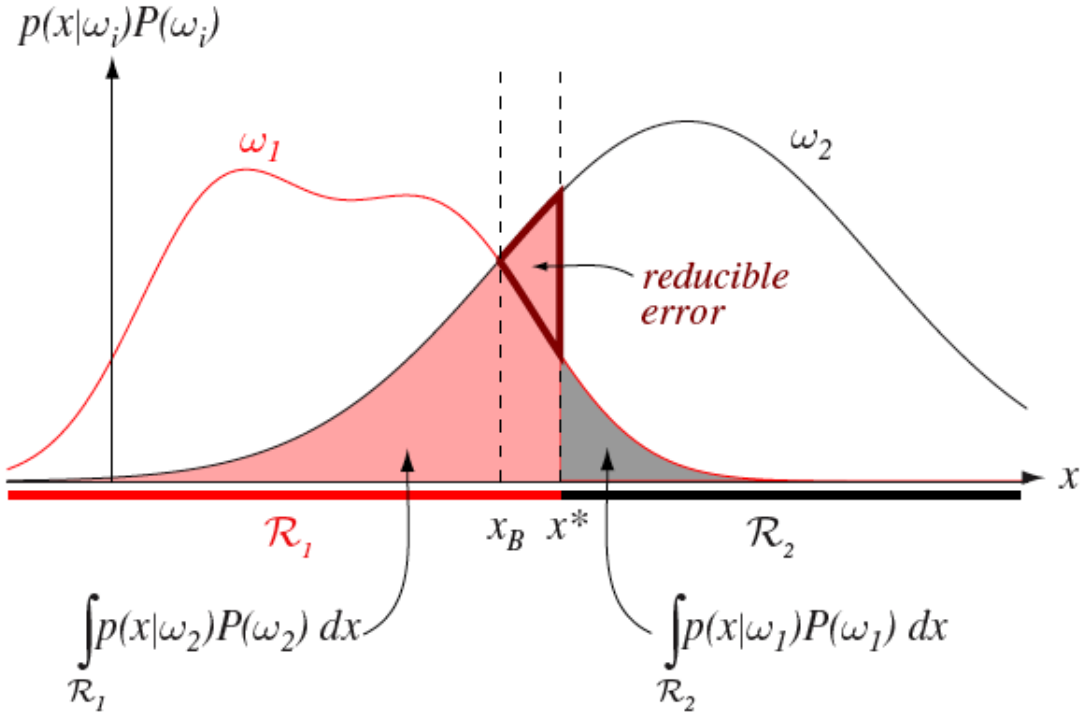
Gleiche A-priori-Wahrscheinlichkeit in allen Klassen führt auf einen linearen Klassifikator (Trennfläche bestimmt sich aus dem kleinsten Abstand zum Klassenmittelwert).

Eigenschaften des naiven Bayes-Klassifikators

- Trotz der stark vereinfachenden Annahmen funktioniert der naive Bayes-Klassifikator in einigen Anwendungen erstaunlich gut, z.B. bei der Dokumentenklassifikation, Gehirnscans und als Spamfilter.
- Man kann theoretisch zeigen, dass die Nachteile der Annahme von unabhängigen Merkmalen bei vielen real vorkommenden Verteilungen nicht sehr ins Gewicht fallen (H. Zhang, Proc. FLAIRS, 2004).
- Die Schätzung der $4 \times d$ Parameter μ_{ij} und σ_{ij} bei d-dimensionalen Daten braucht relative wenig Trainingsdaten. Der Trainingsvorgang ist extrem schnell - $O(\ell)$.
- Je nach Art der Daten gibt es verschiedene Varianten mit unterschiedlichen Annahmen über die Likelihoods der Merkmale: für kontinuierliche Daten GNB, für Binärdaten *Bernoulli Naïve Bayes*, Integerdaten *Multinomial Naïve Bayes*.

Übersicht

Klassifikationsfehler



[Duda et al., 2001]

Es gibt zwei Möglichkeiten für Fehler:

- Beobachtung $\mathbf{x}$ fällt in in $\mathcal{R}_2$, der wahre Naturzustand ist ω_1.
- Beobachtung $\mathbf{x}$ fällt in in $\mathcal{R}_1$, der wahre Naturzustand ist ω_2.

Bayesfehler

Fehlerwahrscheinlichkeit:

$$\begin{aligned} p(\text{Fehler}) &= p(\mathbf{x} \in \mathcal{R}_2, \omega_1) + p(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\ &= p(\mathbf{x} \in \mathcal{R}_2 | \omega_1) p(\omega_1) + p(\mathbf{x} \in \mathcal{R}_1 | \omega_2) p(\omega_2) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_1) p(\omega_1) \, d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_2) p(\omega_2) \, d\mathbf{x} \end{aligned}$$

Der minimale Fehler wird erreicht, wenn die Entscheidungsgrenze auf dem Punkt liegt, an dem beide A-posteriori-Wahrscheinlichkeiten gleich groß sind ($\Rightarrow$ Bayessche Entscheidungsregel).

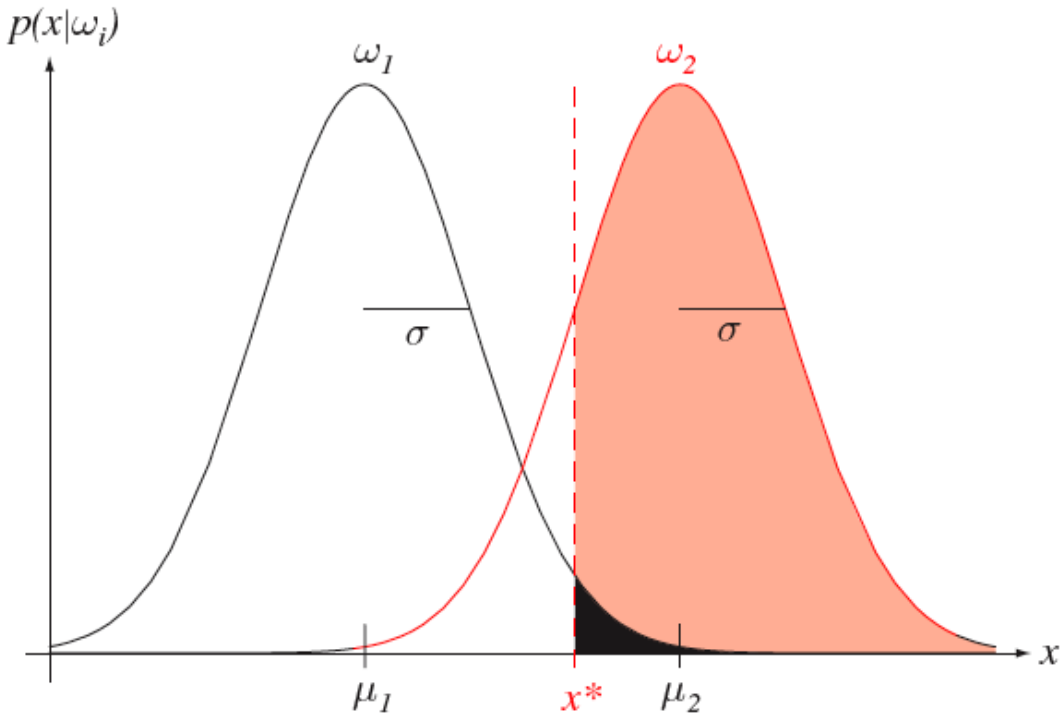
Der dadurch erreichbare minimale Fehler heißt Bayesfehler.

4 mögliche Ausgänge eines Detektionsexperimentes

2 Zustände in der Signaldetektion (Dichotomie): ω_1 - zu detektierendes Signal ist tatsächlich da, ω_2 - Signal nicht vorhanden.

- Hit: Vorhandenes Signal wurde detektiert (auch richtig positiv). Detektionswahrscheinlichkeit (Sensitivität, Richtig-Positiv-Rate): $p(x \in \mathcal{R}_1|\omega_1)$.
- Korrekte Rückweisung: Signal ist nicht vorhanden und wurde auch nicht detektiert (auch richtig negativ). Richtig-Negativ-Rate (Spezifität): $p(x \in \mathcal{R}_2|\omega_2)$.
- Fehlalarm: Detektion trotz nicht vorhandenem Signal (auch falsch positiv, Fehler 1. Art). Fehlalarmrate (Falsch-Positiv-Rate): $p(x \in \mathcal{R}_1|\omega_2)$.
- Miss: Vorhandenes Signal wurde nicht detektiert (auch falsch negativ, Fehler 2. Art). Falsch-Negativ-Rate: $p(x \in \mathcal{R}_2|\omega_1)$.

Beispiel: Signaldetektion bei gaußverteiltem Rauschen

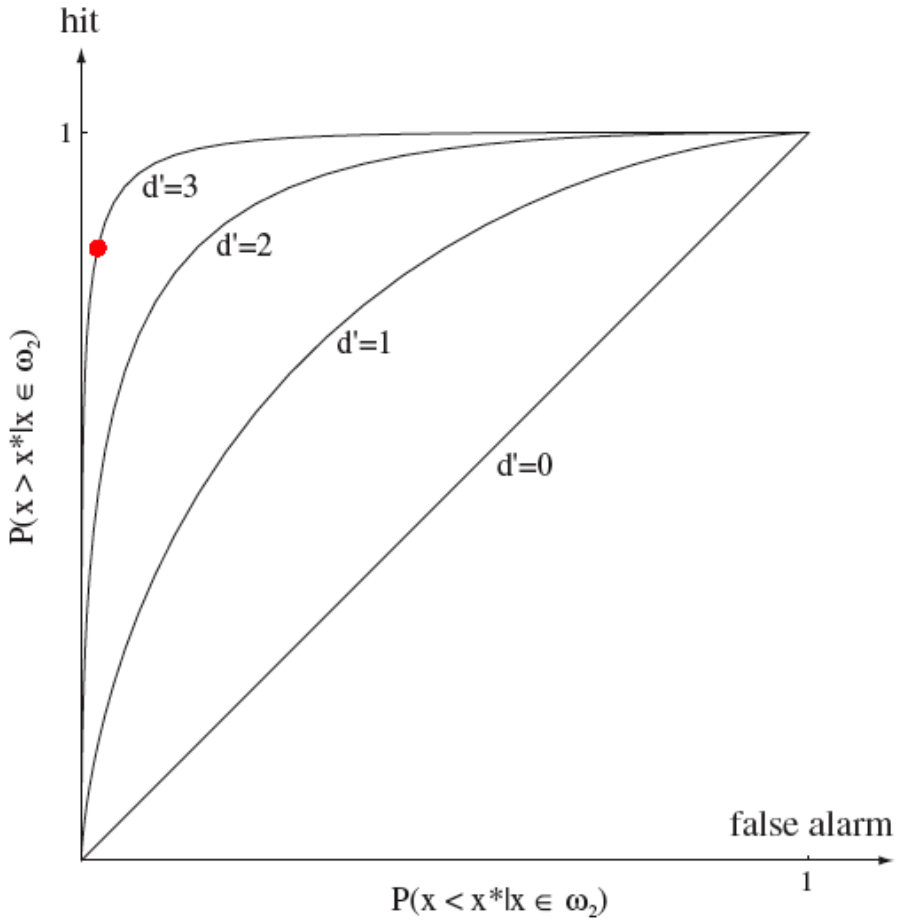


[Duda et al., 2001]

Maß für Unterscheidbarkeit (unabhängig vom Klassifikator):

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}$$

ROC-Kurve

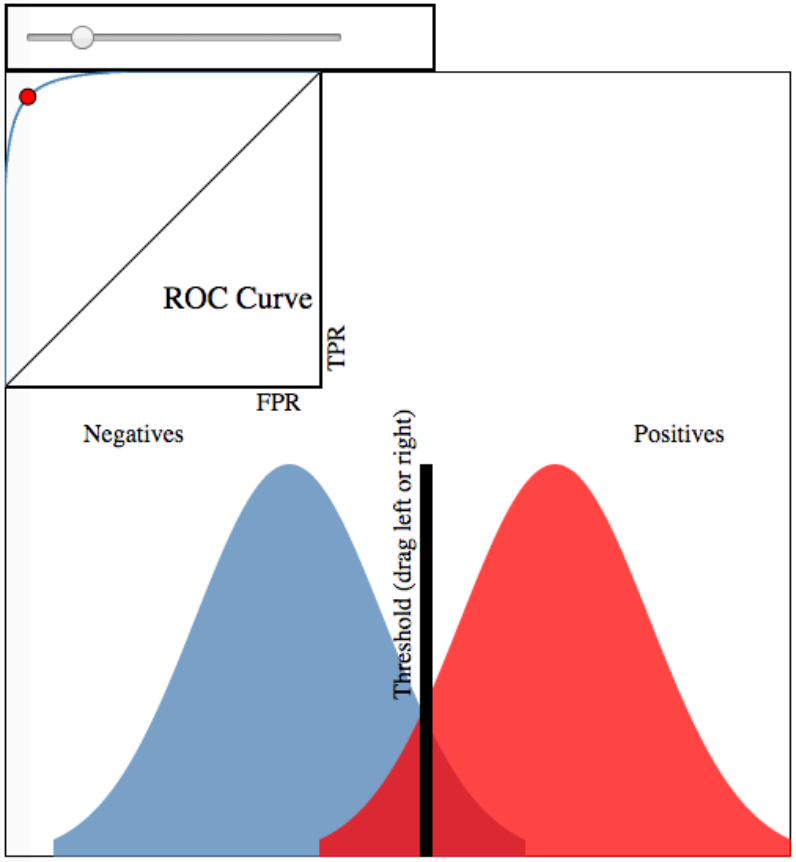


[Duda et al., 2001]

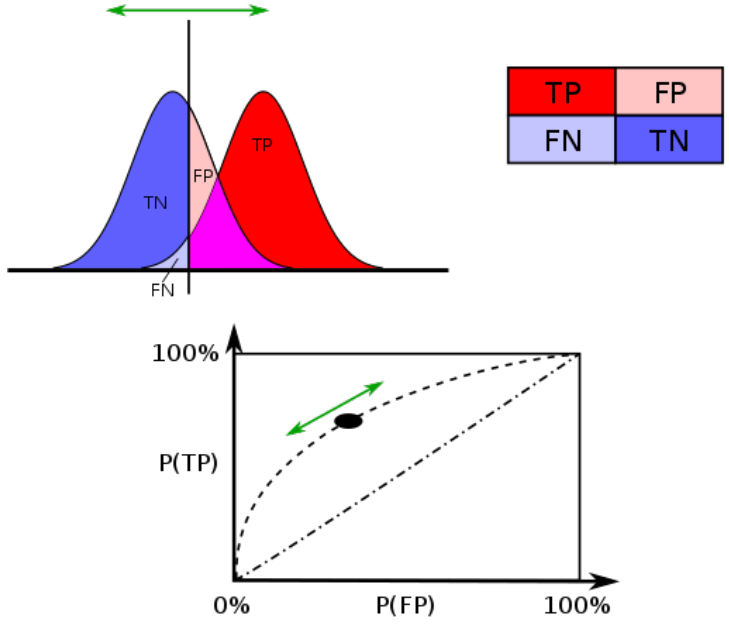
ROC: Receiver Operating Characteristic

Darstellung der Detektions- vs. Fehlalarmrate für alle Schwellwerte

Wie interpretiert man ROC-Kurven?



[http://www.navan.name/roc/]

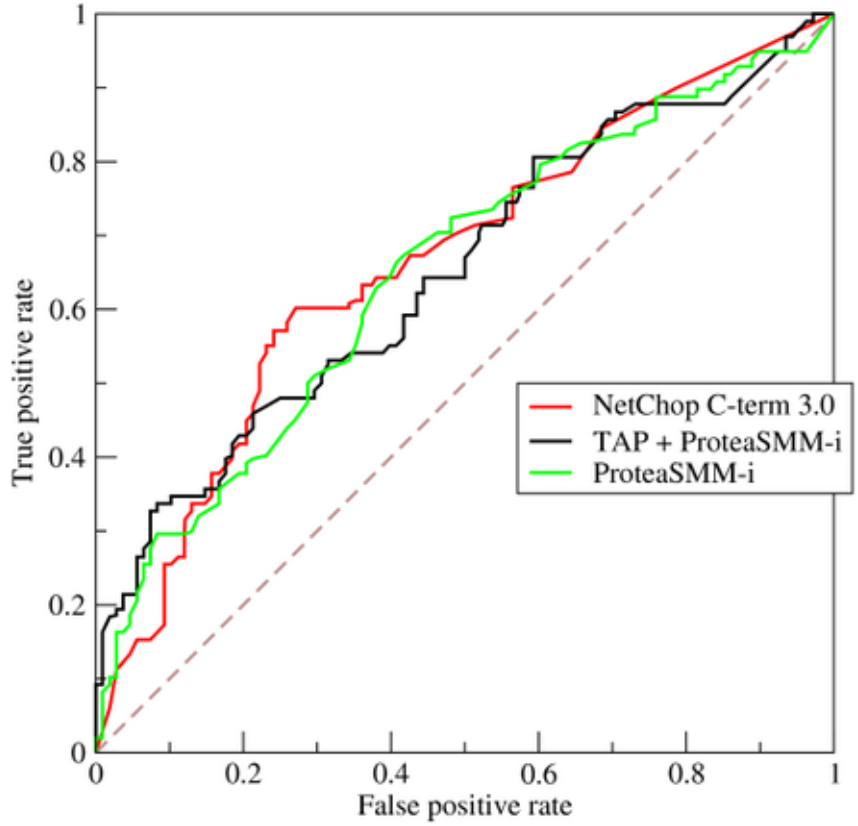


[Wikipedia]

Ausführliches Video:
http://www.dataschool.io/roc-curves-and-auc-explained/

Area Under Curve

- Wird nur eine Detektions- und Fehlalarmrate angegeben, so kann man sich nur mit anderen Klassifikatoren vergleichen, wenn diese die gleiche Fehlalarmrate haben. Vergleiche an unterschiedlichen Punkten der ROC-Kurve sind sinnlos!
- Idealerweise sollte man die volle ROC-Kurve angeben, oder - falls man einen einzigen Kennwert braucht - die Fläche unter der ROC-Kurve (Area Under Curve, AUC). Je höher, desto besser ist der Klassifikator.



[Wikipedia]

- Man kann zeigen, dass der AUC-Wert die Wahrscheinlichkeit angibt, dass der Klassifikator einem positiven Beispiel einen höheren Wert der Entscheidungsfunktion zuweist als einem negativen Beispiel.